Module 2: Assignment 1

Yanyang Pei (B3)

Introduction

As a “superfruit”, blueberries have been proved to be associated with several health benefits, such as maintaining blood sugar levels, prevention of cardiovascular diseases, antimicrobial and antitumoural activity [1]. This has gained interest in the food industry to develop blueberry extract supplements. The nutritional composition of blueberries, particularly their vitamin content, is influenced by a variety of factors including genotype, environmental conditions, and agricultural practices.

Company “Old Nordic” is trying to investigate a vitamin supplement from blueberries extraction. The study primarily aims to understand how vitamin content varies among different types of blueberries grown in various locations and across different years, to provide the company with valuable insights, enabling them to select blueberry varieties that best meet their nutritional goals.

The hypothesis is that the blueberry type is the main factor that determines the vitamin level of blueberries, while growing year doesn’t contribute too much to nutrition. The growing location influences the temperature and rainfall conditions, which in turn affect the vitamin content of blueberries. Blueberries grown in regions with higher temperatures and greater rainfall might have higher vitamin levels.

Materials and Methods

Data analyses were done with Python 6.0. The dataset contained information on four varieties of blueberries—Dixi, Atlantic, Gorgias, and Sweet, which were cultivated across three different years (2020, 2021, and 2022) in five distinct locations: Rockport, Arkham, Dunwich, Ipswich, and Innsmouth. The vitamin content of these blueberries, including levels of Vit B9, Vit K, Vit E, Vit B3, Vit B1, Vit C, and Vit A, was tested.

Grubbs test followed by scatter plot was used to help preprocess the data, and a dendrogram and correlation plot were made to help understand the data.

Studies of the effect of conditions on blueberries nutrient was mainly achieved by using PCA followed by Kruskal-Wallis analysis as well as univariate analysis.

Results

Preprocess and prestudy

Two outliers were found by using Grubbs test, both of them appeared in sample 11. These outliers are illustrated in the scatter plots shown in Figure 1. Together with two blank values, these outliers were replaced with the mean of the corresponding replicates. The adjustments made to the data are summarized in Table 1.

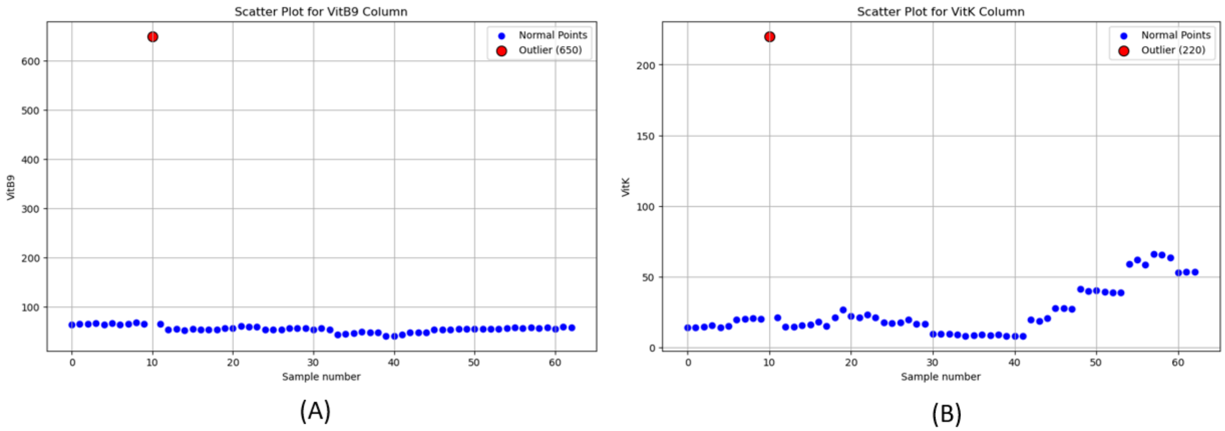


Figure 1. Scatter plot that marks outlier for (A) Vitamin B9, and (B) Vitamin C.

Table 1. Pre-processing of data.

ID	Column	Problem	Found by	Replacement
11	VitB9	Outlier: 650	Grubbs test	65.495
11	VitK	Outlier: 220	Grubbs test	20.7
14	Temp	Empty	Observation	25.72
21	VitB3	Empty	Observation	78.385

By conducting the Shapiro-Wilk test, the normality of the data was studied. All the dependent variables are not normally distributed, output from python is shown in Figure A.1 in Appendix.

The result from cluster analysis shown in Figure 2 indicates that blueberry type Sweet and Dixi are most similar, but least similar to type Atlantic. While the Gorgias group shows more overlap with the clusters of other types, which might suggest that it has a wider range of characteristics or that it shares some characteristics with other types.

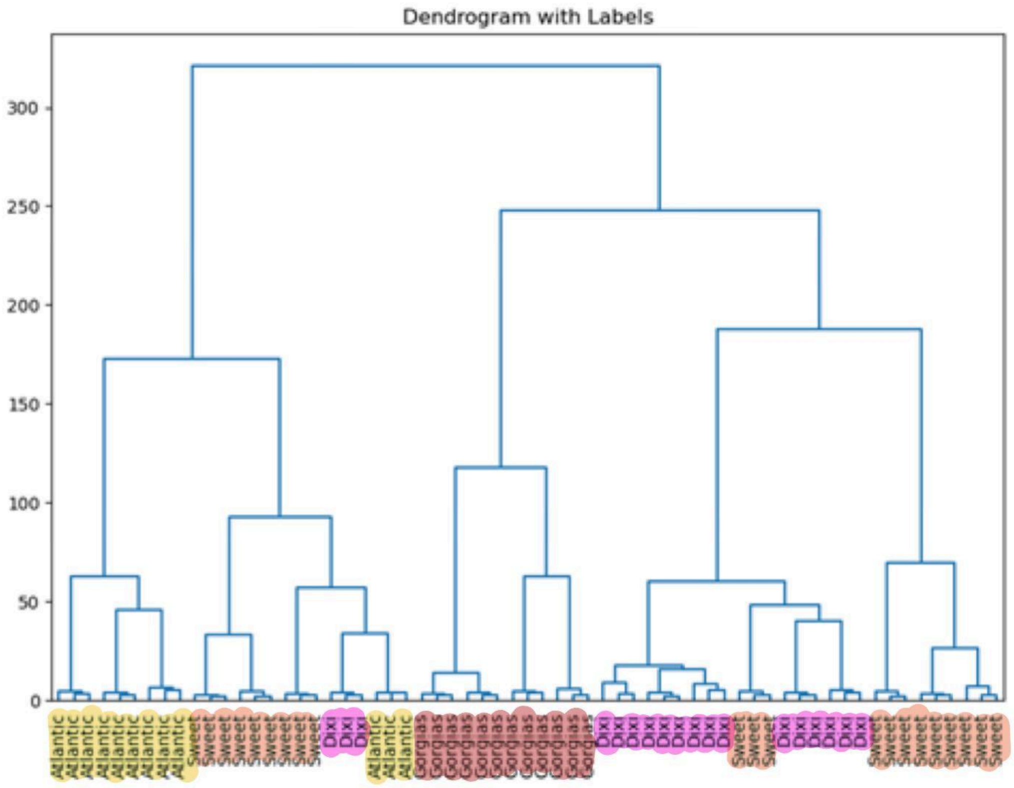


Figure 2. Cluster plot of full data.

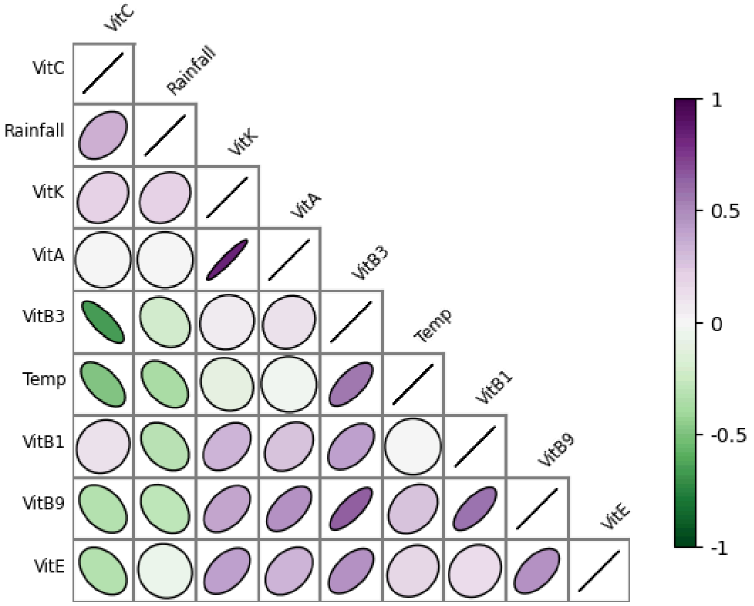


Figure 3. A correlation plot for dependent variables, in which a more elliptical pattern indicates stronger correlation.

The correlation plot in Figure 3 highlights a positive relationship among the three B-group vitamins (B1, B3, and B9) as well as between vitamin A and vitamin K. On the contrary, there is a significant negative correlation between vitamin C and both the B-group vitamins and temperature. Additionally, rainfall appears to have a weak negative correlation with most of the vitamins.

Study of blueberry types

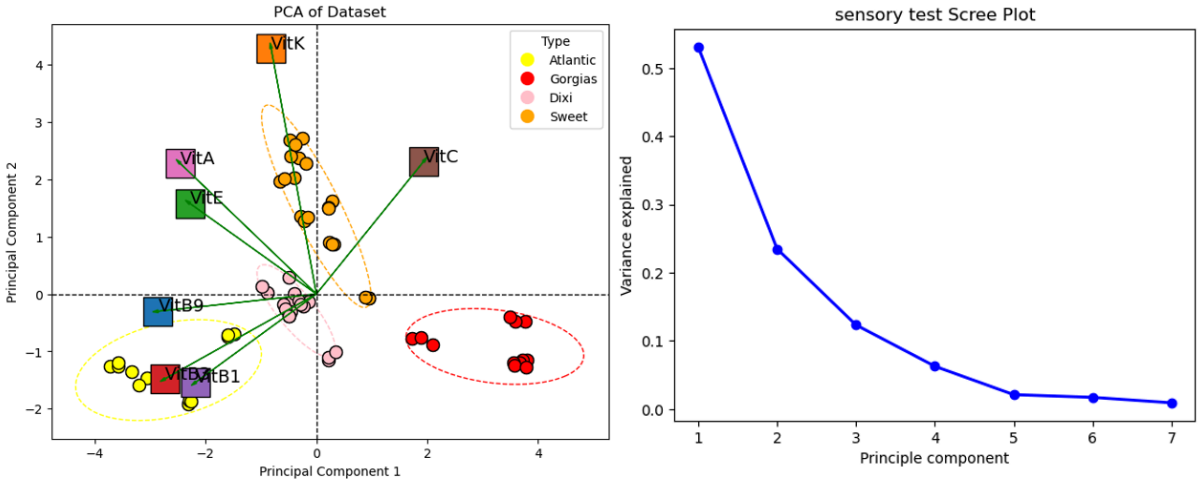


Figure 4. Biplot of scores and loading grouped by blueberry type (left), and scree plot (right).

Figure 4 presents a PCA biplot grouped by blueberry types, with vitamins used as the loading variables. The biplot reveals that the Atlantic blueberry type clusters closely with the B-group vitamins, while the Sweet type is more closely associated with vitamin A and vitamin K. The Gorgias type appears to be more related to vitamin C. Within the loading variables, the B-group vitamins show a strong correlation with each other, suggesting that one of them could be used as a representative for further studies. In this case, vitamin B3 was selected as the representative. Besides, vitamin C and vitamin K show negative correlation with them.

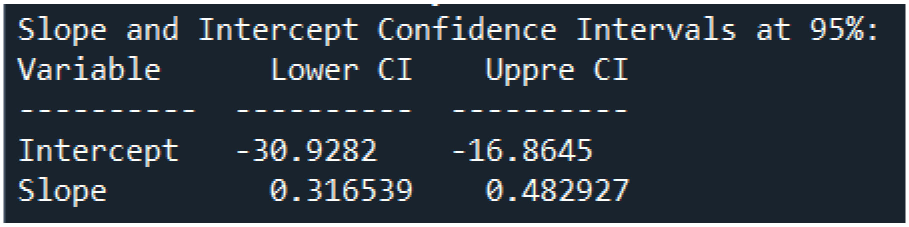
The scree plot indicates that Principal Component 1 (PC1) explains approximately 53% of the variance, while Principal Component 2 (PC2) accounts for about 23%, together accounting for around 76% of the total variability in the dataset. On the PC1 axis, there is a significant distinction between vitamin C and the other vitamins. In contrast, the separation on the PC2 axis primarily differentiates between the other vitamin groups.

The loading of PCA1 and PCA2 are also calculated and shown in Table 2, which helps the discussion of what they represent.

Table 2. Loadings for PCA1 and PCA2:

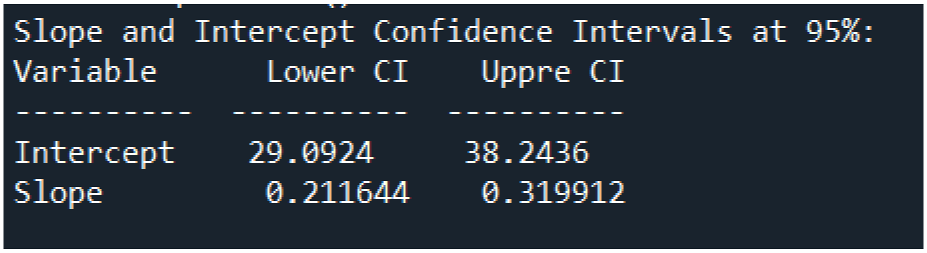
	PCA1	**PCA2**
VitB9	-0.475871	-0.050115
VitK	-0.136001	0.713801
VitE	-0.380364	0.263197
VitB3	-0.455125	-0.246355
VitB1	-0.362718	-0.256193
VitC	0.322099	0.384822
VitA	-0.409784	0.379855

The relationships among the B-group vitamins were examined using linear regression with confidence intervals. The regression between vitamin B1 and B3 yielded an R^2 value of 0.6, while the regression between B9 and B3 resulted in an R^2 of 0.61. These results are illustrated in Figure A.2 in the Appendix. The confidence intervals for the slopes of both regressions (as shown in Figures 5 and 6) do not include zero, indicating a statistically significant relationship.



```
Slope and Intercept Confidence Intervals at 95%:
Variable      Lower CI    Uppre CI
----------  ----------  ----------
Intercept   -30.9282    -16.8645
Slope         0.316539    0.482927
```

Figure 5. Confidence intervals for regression between B1 and B3



```
Slope and Intercept Confidence Intervals at 95%:
Variable      Lower CI    Uppre CI
----------  ----------  ----------
Intercept    29.0924     38.2436
Slope         0.211644    0.319912
```

Figure 6. Confidence intervals for regression between B9 and B3

Table 3 provides a summary of the significance of differences in vitamin levels across various blueberry types, along with the specific blueberry type that exhibits the highest level for each vitamin. It can be concluded that blueberry types cause significant differences in vitamin levels. Barchart for different levels are shown in Figure A.3 in Appendix.

Table 3. Summary of effect on blueberry type on vitamin

	Vitamin	**Kruskal Wallis P value**	**Significance**	**Highest (from barchart)**	**Mann-Whitney p value**	**Significance**
Type	B3 (represent B)	1.2e-10	**	Atlantic	0.000	**
	C	2.1e-10	**	Gorgias	5.6e-09	**
	E	9.1e-06	**	Atlantic	0.819	ns
	A	3.5e-08	**	Sweet	0.001	**
	K	2.1e-10	**	Sweet	8.7e-08	**

ns: P- value >0.05, “”: P- value <=0.05 and “**”: P- value <=0.01

Study of growing year

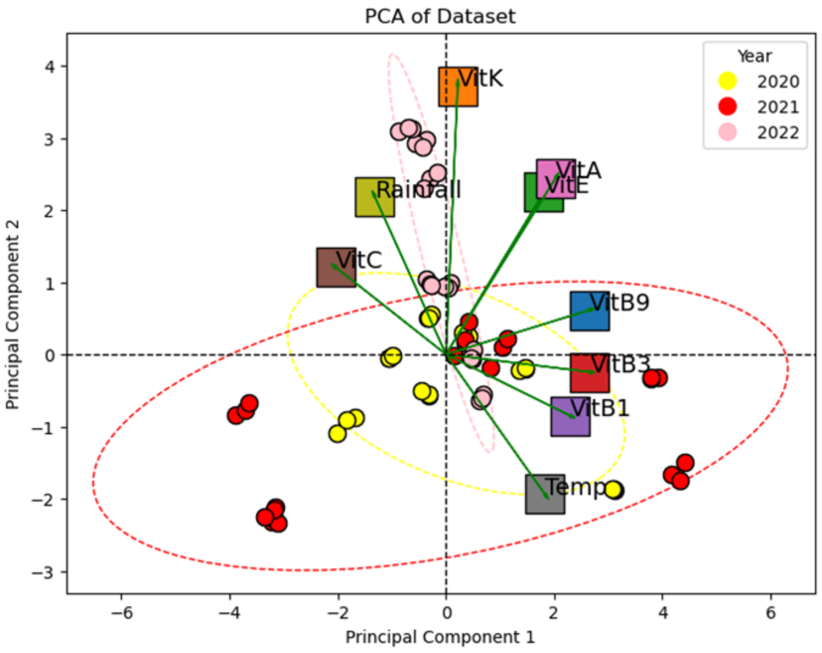


Figure 6. Biplot with scores and loading, grouped by growing year.

The PCA shown in Figure 6 reveals that most of the scores data are quite spread out, particularly for the year 2021. This suggests significant variability within each year, indicating that growing years alone may not explain the differences well.

Table 4 summarizes the significance of differences in vitamin levels across the different growing years, along with the specific year that exhibits the highest level for each vitamin. It can be seen that most of the vitamins are not strongly affected by growing year, except vitamin A and vitamin K. Additionally, the differences observed among the B-group vitamins suggest that using B3 as a representative may result in a loss of important information. The bar chart comparing vitamins grown in different years are shown in Figure A.4 in Appendix.

Table 4. Summary of effect on growing year on vitamin.

	Vitamin	**Kruskal Wallis P value**	**Significance**	**Highest (from barchart)**	**Mann-Whitney p value**	**Significance**
Year	B3	0.033	*	2020	0.126	ns
	B1	0.111	ns			
	B9	0.628	ns			
	C	0.147	ns			
	E	0.596	ns			
	A	1.105e-06	**	2022	1.825e-06	**
	K	9.297e-06	**	2022	6.272e-07	**

Studying of growing place:

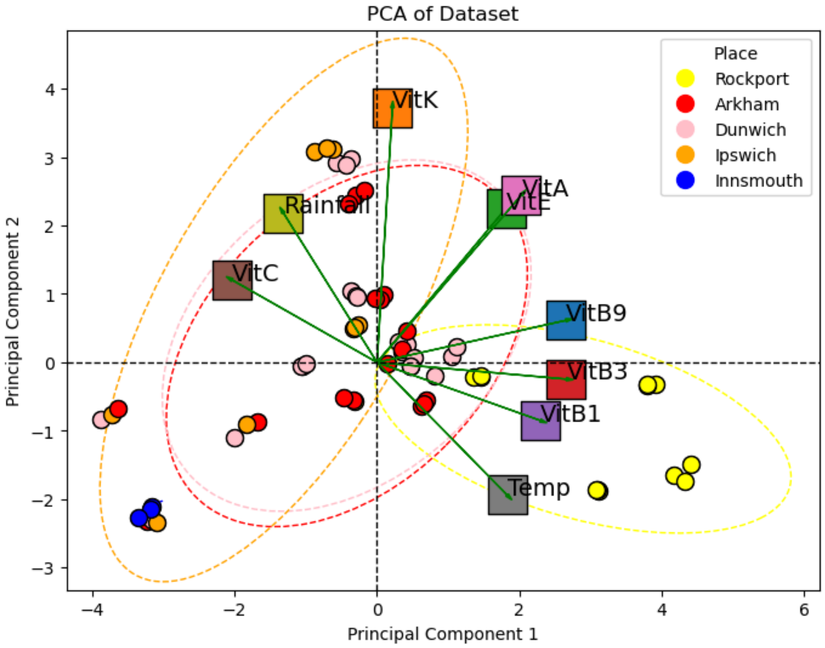


Figure 7. Biplot with scores and loading, grouped by growing place.

In this PCA shown in Figure 7, the blueberry grown in Rockport seems to cluster closely with vitamin B group, indicating it might have higher concentration of these vitamins. The blueberries from Arkham and Dunwich are widely spread but tend to be in the center, suggesting lots of variability but overall average amount of vitamins without extreme values. Besides, points from Ipswich seem to be closer towards the vitamin K area, while Innsmouth only has very few points and doesn't explain a lot.

Table 5 summarizes the significance of differences in vitamin levels across the different growing places, along with the specific place that exhibits the highest level for each vitamin. The data indicate that vitamins A and K are not significantly influenced by the growing location. Overall, Rockport is the location that produces the most nutritious blueberries. The bar chart comparing vitamins in different places are shown in Figure A.5 in Appendix.

Table 5. Summary of the effect on growing places on vitamins.

	Vitamin	**Kruskal Wallis P value**	**Significance**	**Highest (from barchart)**	**Mann-Whitney p value**	**Significance**
Place	B3 (represent B)	1.767e-06	**	Rockport	8.03e-07	**

	C	0.001	**	Ipswich	0.009	**
	E	0.027	*	Rockport	0.506	ns
	A	0.106	ns			
	K	0.068	ns			

Study of temperature and rainfall

From PCA shown in Figure 5, it can be seen that temperature clusters more closely with the year 2020, while rainfall is more aligned with the year 2022, suggesting a potential correlation between these environmental factors and the respective years. In contrast, 2021 appears to have experienced lower levels of rainfall and temperature. And it's noticeable that in Figure 6, area Rockport tends to have higher temperatures than other places, while Ipswich seems to have higher rainfall. By combining both bioplots, it can be concluded that temperature has strong positive correlation with vitamin B1 and B3 and B9, but negative correlation with vitamin C. On the contrary, rainfall has positive correlation with vitamin C and K, but negative correlation with B-group vitamins. This aligns with the result in the correlation plot in Figure 3.

In addition, the relationship between place and year with these two environmental factors were studied by using Kruskal-Wallis test along with Mann-Whitney as post hoc test, the result was summarized in Table 6. It reveals that both year and place have significant impact on temperature, but rainfall is only dependent on place. The bar chart comparing temperatures and rainfall in different places and year are shown in Figure A.6 in Appendix.

Table 6. Kruskal-Wallis statistic for testing the effect of place and year on temperature and rainfall.

		Kruskal Wallis P value	**Significance**	**Highest (from barchart)**	**Mann-Whitney p value**	**Significance**
Place	Temp	0.033	*	Rockport	0.070	ns
	Rainfall	0.000	**	Ipswich	0.624	ns
Year	Temp	0.012	*	2021	0.145	*
	Rainfall	0.420	ns			

Discussion

The whole study proves the hypothesis that blueberry types impact most of the differences in nutrients and the growing year almost has no effects on vitamins. However the hypothesis was wrong that temperature and rainfall is not positively correlated with all the vitamin content. The PCA analysis and univariate study successfully support the relationship provided in the correlation plot.

The two principal components in the PCA plot, using Figure 4 as an example, can be interpreted through the loadings. PC1, which shows a clear separation between vitamin C and the other vitamins, effectively represents the contrast between samples with higher levels of VitC and those with higher levels of the other vitamins. The loading data presented in Table 2 supports this interpretation: vitamins B9, B3, B1, A, and E exhibit strong negative loadings, significantly influencing sample scores, while vitamin C has a strong positive loading. Consequently, samples with higher vitamin C levels will have higher PC1 scores. Although vitamin K also has a negative loading, it is relatively low compared to the other vitamins, indicating a lesser effect on PC1. As for PC2, there's a good separation between B-group vitamins and others, representing the contrast between these groups. The loading data reveals that vitamin K has a very high positive loading, suggesting that PC2 scores are mainly driven by level of vitamin K. Additionally, vitamins C, A, and E also contribute to positive PC2 scores. In contrast, all B-group vitamins exhibit negative loadings, with vitamin B9 showing the weakest influence among them.

It can be concluded that, to gain higher vitamin B-group content, it's suggested to choose blueberries type Atlantic from Rockport, with high temperature growing conditions. On the other hand, if providing vitamin C content is the main goal, blueberries type Gorgias from Ipswich, with heavy rainfall growing condition can be selected. Another study on wheat grain also suggests that higher temperature provides a higher amount of vitamin B in crops [2].

Issue with experiment data and suggested further research

In the data, there's observation that within the **Gorgias** group, temperature and rainfall vary significantly within the same year and place (Figure A.7 in Appendix), while other groups (such as Atlantic, Dixi and Sweet) show similar temperature and rainfall within the same year and place, which suggests there might be error in data recording. Hence, it's recommended to review data or do focused study to investigate the cause.

Also, in the same blueberry type group, which was grown in Ipswich in the year 2021, there's a large difference in the vitamin K level. However, since only two data is presented, the outlier can be determined. It's therefore suggested to reconduct the experiment on this group.

Besides, given that some samples in this experiment, such as those from the Innsmouth and Ipswich areas, have a limited number of replicates, increasing the number of experiments could enhance the accuracy and reliability of the results.

Future research should explore the mechanism of how temperature and rainfall affect vitamin synthesis in blueberries. Additionally, expanding this analysis to other berries, and exploring the effects of other environmental variables, such as soil composition and sunlight exposure, could provide a more comprehensive understanding of how to optimize agricultural practices for nutrient-rich produce.

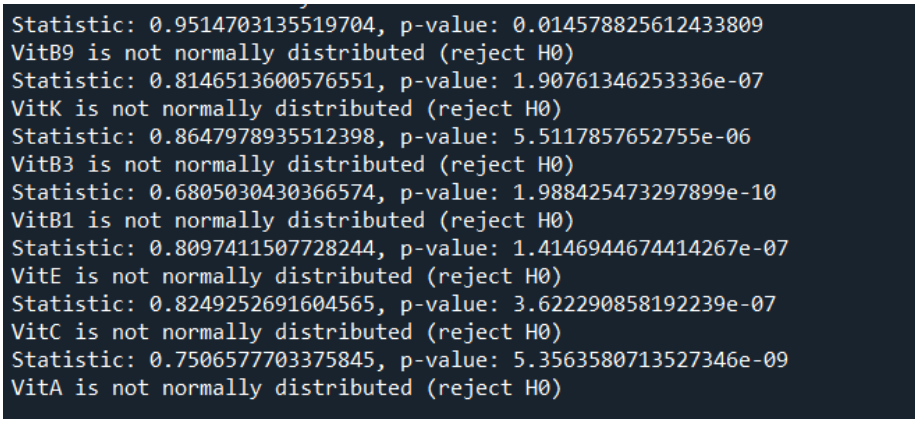
Conclusion

The nutritional content of blueberries was analyzed using PCA and univariate analysis. The results indicate that blueberry type is the primary factor influencing vitamin content. The Atlantic variety produces the highest levels of B-group vitamins and vitamin E, while the Sweet variety is richest in vitamins A and K. The Gorgias variety, on the other hand, contains the highest levels of vitamin C.

In addition to the significant impact of blueberry type, environmental conditions were also found to play a crucial role in determining vitamin content. The analysis shows that higher temperatures are associated with increased B-group vitamin content, while heavier rainfall is correlated with higher vitamin C levels. Consequently, Rockport, which experiences higher temperatures, has the highest levels of B-group vitamins, while Ipswich, known for its heavy rainfall, shows the highest levels of vitamin C.

Interestingly, growing year overall doesn’t have a strong impact on blueberry vitamin amount.

Appendix



```
Statistic: 0.9514703135519704, p-value: 0.014578825612433809
VitB9 is not normally distributed (reject H0)
Statistic: 0.8146513600576551, p-value: 1.90761346253336e-07
VitK is not normally distributed (reject H0)
Statistic: 0.8647978935512398, p-value: 5.5117857652755e-06
VitB3 is not normally distributed (reject H0)
Statistic: 0.6805030430366574, p-value: 1.988425473297899e-10
VitB1 is not normally distributed (reject H0)
Statistic: 0.8097411507728244, p-value: 1.4146944674414267e-07
VitE is not normally distributed (reject H0)
Statistic: 0.8249252691604565, p-value: 3.622290858192239e-07
VitC is not normally distributed (reject H0)
Statistic: 0.7506577703375845, p-value: 5.3563580713527346e-09
VitA is not normally distributed (reject H0)
```

Figure A.1 Python output for Shapiro-Wilk test.

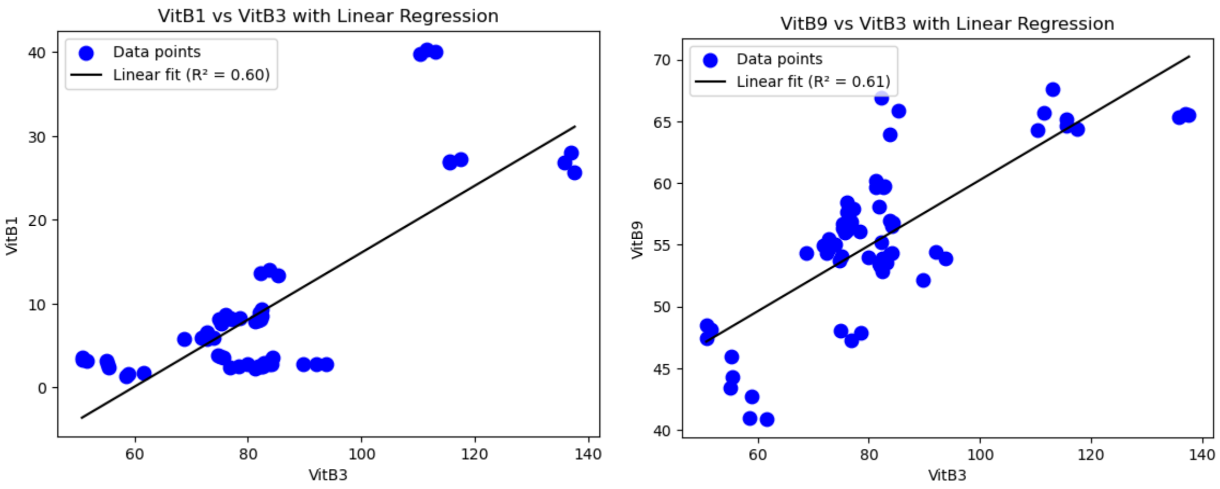


Figure A.2. Linear regression plot between B-group vitamins.

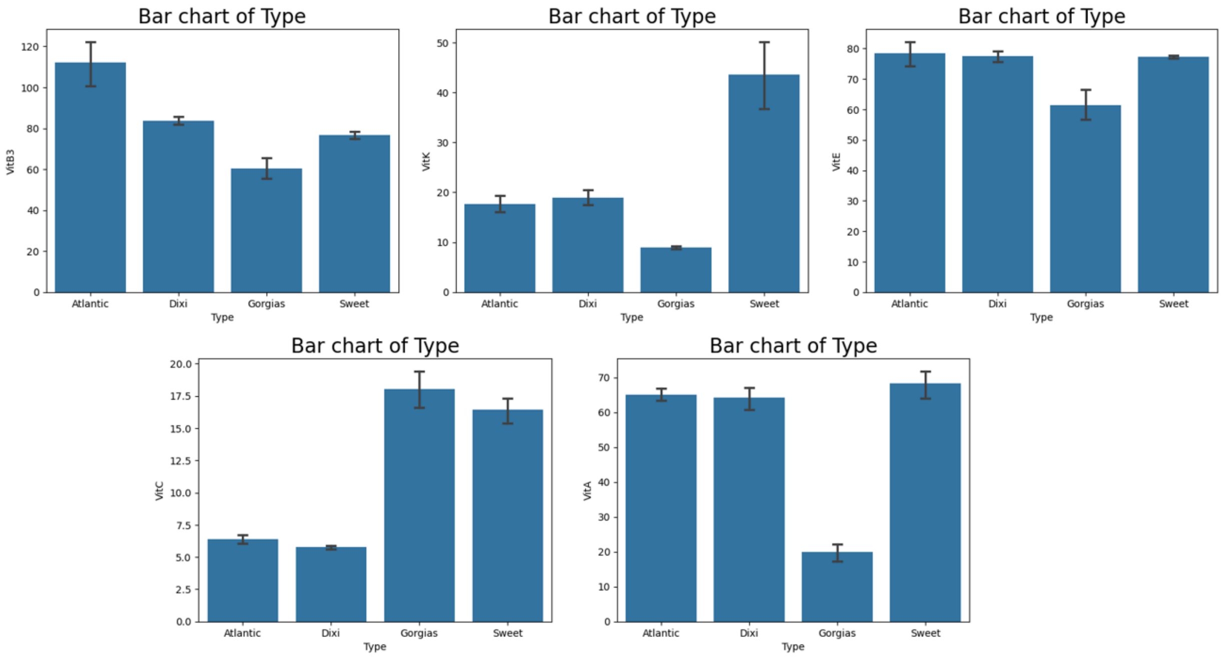


Figure A.3. Barchart showing vitamins differences within variant blueberry types.

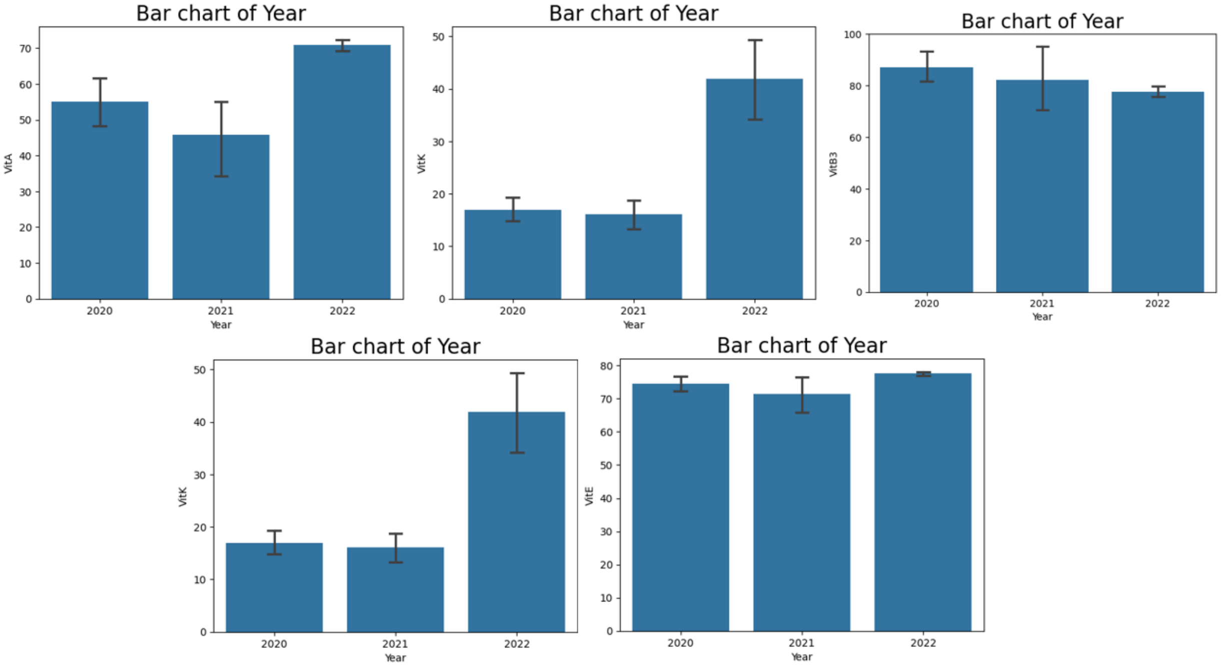


Figure A.4. Barchart showing vitamins differences within variant growing years.

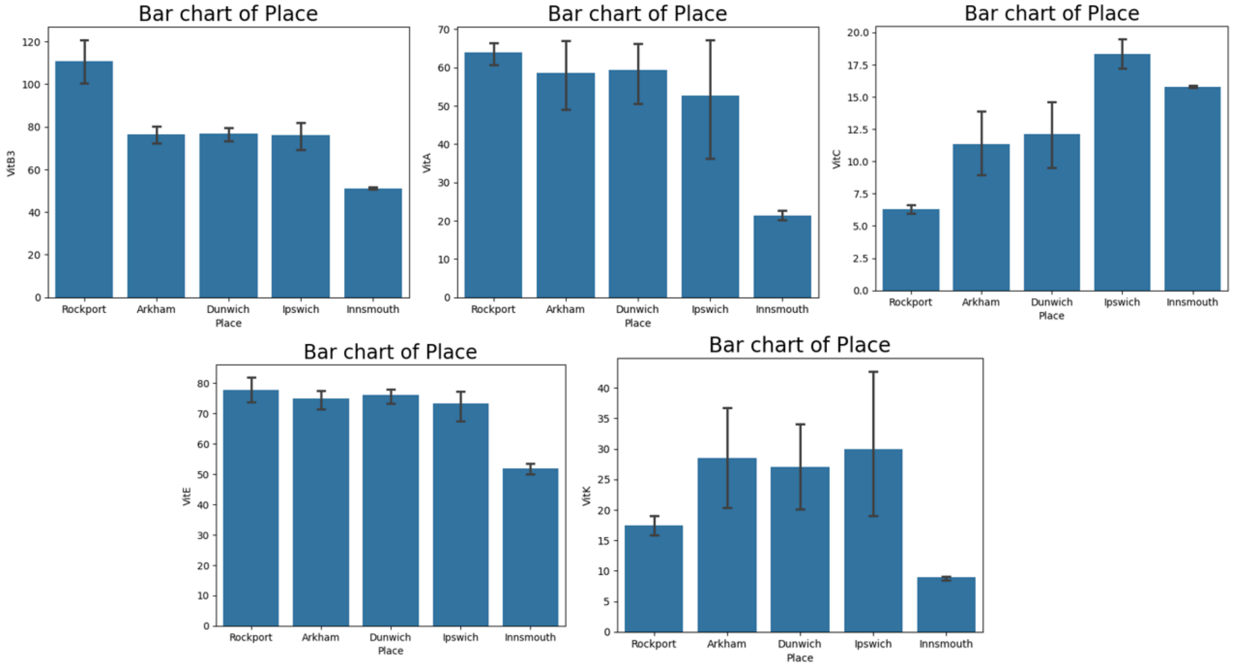


Figure A.5. Barchart showing vitamins differences within variant growing places.

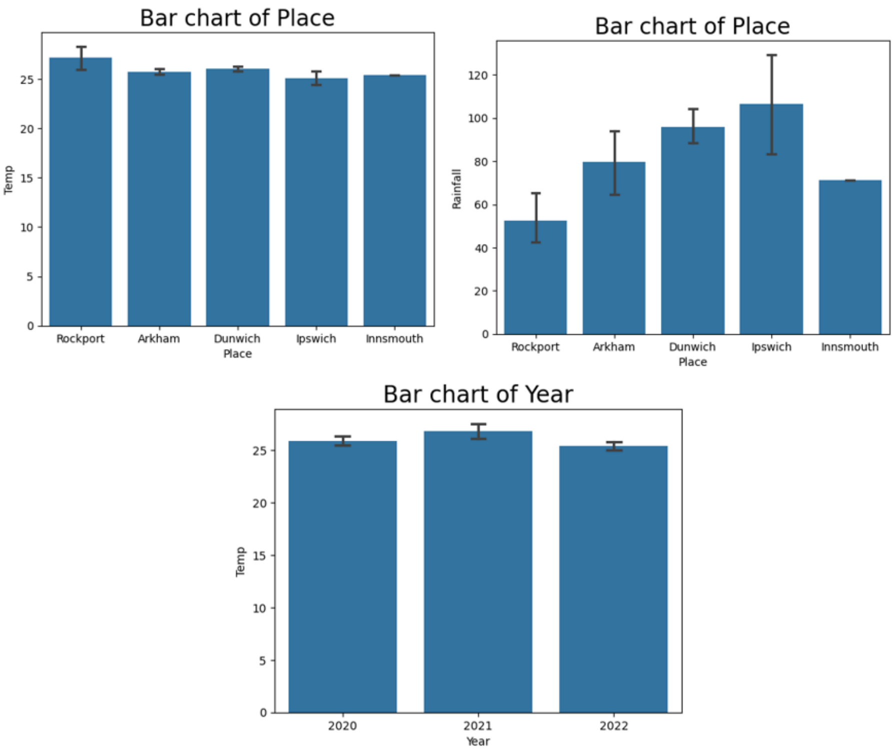
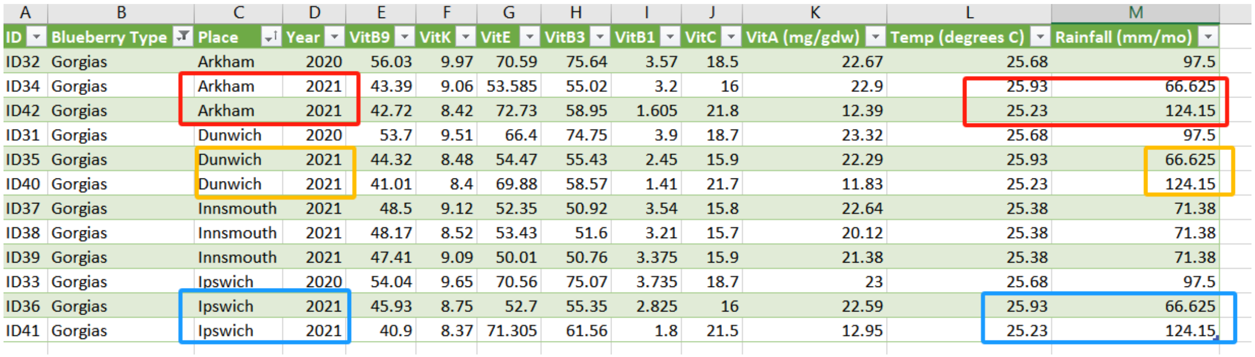


Figure A.6. Barchart showing temperature and rainfall differences within variant place and year.



A	B	C	D	E	F	G	H	I	J	K	L	M
ID	Blueberry Type	Place	Year	VitB9	VitK	VitE	VitB3	VitB1	VitC	VitA (mg/gdw)	Temp (degrees C)	Rainfall (mm/mo)
ID32	Gorgias	Arkham	2020	56.03	9.97	70.59	75.64	3.57	18.5	22.67	25.68	97.5
ID34	Gorgias	Arkham	2021	43.39	9.06	53.585	55.02	3.2	16	22.9	25.93	66.625
ID42	Gorgias	Arkham	2021	42.72	8.42	72.73	58.95	1.605	21.8	12.39	25.23	124.15
ID31	Gorgias	Dunwich	2020	53.7	9.51	66.4	74.75	3.9	18.7	23.32	25.68	97.5
ID35	Gorgias	Dunwich	2021	44.32	8.48	54.47	55.43	2.45	15.9	22.29	25.93	66.625
ID40	Gorgias	Dunwich	2021	41.01	8.4	69.88	58.57	1.41	21.7	11.83	25.23	124.15
ID37	Gorgias	Innsmouth	2021	48.5	9.12	52.35	50.92	3.54	15.8	22.64	25.38	71.38
ID38	Gorgias	Innsmouth	2021	48.17	8.52	53.43	51.6	3.21	15.7	20.12	25.38	71.38
ID39	Gorgias	Innsmouth	2021	47.41	9.09	50.01	50.76	3.375	15.9	21.38	25.38	71.38
ID33	Gorgias	Ipswich	2020	54.04	9.65	70.56	75.07	3.735	18.7	23	25.68	97.5
ID36	Gorgias	Ipswich	2021	45.93	8.75	52.7	55.35	2.825	16	22.59	25.93	66.625
ID41	Gorgias	Ipswich	2021	40.9	8.37	71.305	61.56	1.8	21.5	12.95	25.23	124.15

Figure A.7. Odd results in rainfall and temperature.

Reference

[1] Silva, S., Costa, E. M., Veiga, M., Morais, R. M., Calhau, C., & Pintado, M. (2020). Health promoting properties of blueberries: A review. Critical reviews in food science and nutrition, 60(2), 181-200.

[2] Shewry, P. R., Van Schaik, F., Ravel, C., Charmet, G., Rakszegi, M., Bedo, Z., & Ward, J. L. (2011). Genotype and environment effects on the contents of vitamins B1, B2, B3, and B6 in wheat grain. Journal of agricultural and food chemistry, 59(19), 10564-10571.

Acknowledgements

AI was used in code writing as well as report language improvement.